1

Technical Report 923

A Review of and Recommendations for Procedures Used to Evaluate the External Effectiveness of Intelligent Tutoring Systems

Peter J. Legree and Philip D. Gillis
U.S. Army Research Institute

March 1991

United States Army Research Institute
for the Behavioral and Social Sciences

Approved for public release; distribution is unlimited.

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

JON W. BLADES
COL, IN
Commanding

Technical review by

V. Melissa Holland
William J. York, Jr.

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POX, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Technical Report 923

A Review of and Recommendations for Procedures Used to Evaluate the External Effectiveness of Intelligent Tutoring Systems

Peter J. Legree and Philip D. Gillis
U.S. Army Research Institute

Field Unit at Fort Gordon, Georgia
Michael G. Sanders, Chief

Training Research Laboratory
Jack H. Hiller, Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

March 1991

Army Project Number
2Q263007A795

Training Simulation

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS
--

2a. SECURITY CLASSIFICATION AUTHORITY
--

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE
--

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release;
distribution is unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
ARI Technical Report 923

5. MONITORING ORGANIZATION REPORT NUMBER(S)
--

6a. NAME OF PERFORMING ORGANIZATION
U.S. Army Research Institute
Fort Gordon Field Unit

6b. OFFICE SYMBOL *(If applicable)*
PERI-IG

7a. NAME OF MONITORING ORGANIZATION
--

6c. ADDRESS *(City, State, and ZIP Code)*
Building 41203
Fort Gordon, GA 30905-5230

7b. ADDRESS *(City, State, and ZIP Code)*
--

8a. NAME OF FUNDING / SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences

8b. OFFICE SYMBOL *(If applicable)*
PERI-I

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
--

8c. ADDRESS *(City, State, and ZIP Code)*
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
63007A	795	3312	H01

11. TITLE *(Include Security Classification)*
A Review of and Recommendations for Procedures Used to Evaluate the External Effectiveness of Intelligent Tutoring Systems.

12. PERSONAL AUTHOR(S)
Legree, Peter J.; and Gillis, Philip D.

13a. TYPE OF REPORT
Interim

13b. TIME COVERED
FROM 90/03 TO 91/03

14. DATE OF REPORT *(Year, Month, Day)*
1991, March

15. PAGE COUNT

16. SUPPLEMENTARY NOTATION
--

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

Artificial intelligence (AI)	Training
Intelligent tutoring system (ITS)	Education
Computer-based instruction (CBI)	Evaluation

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This report reviews and discusses intelligent tutoring system (ITS) evaluation standards and procedures. Three criteria evolved from the review and are proposed to accurately evaluate ITS product effectiveness. First, the instructional effectiveness of ITS applications, human tutors, and traditional methods needs to be compared using performance data. Second, extensive ITS applications should be used to evaluate instruction effectiveness. Third, large groups of subjects must be used for the evaluations to precisely estimate ITS effectiveness.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Peter J. Legree

22b. TELEPHONE *(Include Area Code)*
(404) 791-5523

22c. OFFICE SYMBOL
PERI-IG

Previous editions are obsolete.

FOREWORD

This report reviews methods used to evaluate intelligent tutoring system technology and makes recommendations for improving intelligent tutoring system evaluation methods. These recommendations will be applied to the evaluation of intelligent tutoring systems at the Signal School. The development and evaluation of intelligent tutoring systems was sponsored by the Signal Corps based on the concern that existing Signal training approaches are not adequate to support future Signal training requirements.

This report documents work that is part of a larger research program established to support the Signal Center and Signal School, U.S. Army Research Institute for the Behavioral and Social Sciences Research Task 3303, "Technologies for Communications and Electronic Skills Training." This project was briefed to the Commanding General of the Signal School and Corps on December 27, 1990.

Edgar M. Johnson

EDGAR M. JOHNSON
Technical Director

A REVIEW OF AND RECOMMENDATIONS FOR PROCEDURES USED TO EVALUATE THE EXTERNAL EFFECTIVENESS OF INTELLIGENT TUTORING SYSTEMS

EXECUTIVE SUMMARY

Requirement:

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) Fort Gordon Field Unit was asked to develop and evaluate intelligent tutoring system (ITS) applications to support Signal training.

Procedure:

The literature was searched to identify ITS evaluation studies. The reports were analyzed to identify problems and promising aspects of the evaluations.

Findings:

The literature review established a need for evaluation research to estimate the product effectiveness of ITS technology. The review indicates a requirement for three conditions in ITS evaluation studies: the ITS group, a traditional instruction group, and a human tutor group. The literature review indicates that large sample sizes should be used in evaluation projects and that only extensive systems should be evaluated.

Utilization of Findings:

This document is being used as the basis for evaluating the following intelligent tutoring systems: Advanced Learning for Mobile Subscriber Equipment, Posit, the Mobile Subscriber Radio-Telephone Terminal Tutor, and the Network Management Facility Tutor.

A REVIEW OF AND RECOMMENDATIONS FOR PROCEDURES USED TO EVALUATE THE EXTERNAL EFFECTIVENESS OF INTELLIGENT TUTORING SYSTEMS

CONTENTS

Page

INTRODUCTION 1

BACKGROUND 1

PAST ITS EVALUATIONS 2

Lisp Tutor 2
Proust 4
West 7
MACH III 8
Smithtown 9
Prewriting Tutor 10
Pixie 10
Adaptive Computerized Training System 12

ITS EVALUATION CRITERIA 14

Required Conditions: ITS, Human Tutor, and Traditional Instruction 14
Impact of the System 16
Sample Size Requirement: Power Analyses 18
Other Considerations 20

SUMMARY 21

REFERENCES 23

LIST OF TABLES

Table 1. Proust evaluation design 5

2. Sample sizes used in ITS evaluations 18

3. Power analysis estimates: Required sample size per group at the alpha = .05 (2-tail) level by effect size and power 19

A REVIEW OF AND RECOMMENDATIONS FOR PROCEDURES USED TO EVALUATE THE EXTERNAL EFFECTIVENESS OF INTELLIGENT TUTORING SYSTEMS

Introduction

The increasing power and decreasing costs of microcomputer hardware and software allow the economical production and distribution of intelligent tutoring system (ITS) applications. Determining the training effectiveness of ITS is critical to the advancement of this technology as a practical instructional delivery system. This paper proposes product effectiveness criteria in order to accurately estimate ITS training effectiveness.

This review is limited to procedures required to evaluate the instructional (external) effectiveness of ITS technology, as opposed to other training technologies. A related issue is the discussion of procedures designed to evaluate the internal effectiveness of an ITS; this issue is addressed by Priest and Young (1988), and Orey, Legree, Gillis, and Bloom (1991). Although internal effectiveness criteria are central to optimizing ITS effectiveness during development, they are not discussed because they provide scant justification for funding real-world ITS application development.

Background

Although over twenty ITS applications have been developed, very few of these systems have been formally evaluated. Many of these systems were developed on Artificial Intelligence workstations that cannot be used to distribute training economically. These systems were primarily designed to demonstrate the types of interactions and approaches that could be used to deliver training. The fact that these ITS applications were not intended for widespread use may account for the paucity of formal evaluations.

ITS evaluations have often been based either on user acceptance data or on content analyses. Although both types of analyses may reveal shortcomings in existing systems and may point to possible improvements, neither type of analysis allows the effectiveness of a tutoring system to be estimated.

Many current ITS applications will run on microcomputers because of improvements in personal computer-based software. These systems have the potential to become cost-effective training systems because they run on readily available microcomputers. However, because ITS applications remain expensive to develop, it is critical that performance data be collected to determine ITS effectiveness.

In the past, the developmental cost of ITS technology has been justified by the claim that this technology may approach (Anderson, Boyle & Reiser, 1985), or surpass (Littman & Soloway, 1988) the effectiveness of one-on-one human tutorial instruction. Reaching this goal would have major implications for training and education because students who are individually taught by human tutors perform approximately two standard deviations better on instructional exams than students taught by traditional classroom methods (Bloom, 1984). The demonstration that ITS technology is highly effective would also justify continued support for developing practical ITS applications.

Evaluating ITS technology with well designed research studies is critical to avoid repeating the errors seen in evaluations of conventional computer-aided instruction (CAI). Initial meta-reviews of CAI studies (Kulik, Bangert & Williams, 1983; Bangert-Downs, Kulik & Kulik, 1985; Kulik, Kulik & Shwalb, 1986) supported optimistic estimates of the effectiveness of CAI and reported effect size estimates ranging from 0.25 to 0.42 standard deviations. However, a meta-analysis (Clark, 1985) that eliminated poorly designed studies from Kulik and Kulik's reviews estimated a much smaller effect size for CAI effectiveness, 0.09 standard deviation. This latter estimate underscores the importance of using sound evaluation procedures to estimate the instructional effectiveness of ITS.

The next section of this paper will review ITS evaluation studies and identify promising approaches. General evaluation guidelines are presented in the final section of the paper.

Past ITS Evaluations

The following ITS evaluations were identified through computer searches and by reviewing frequently cited documents. Each evaluation study is summarized and analyzed as an approach to estimating the instructional effectiveness of ITS technology.

Lisp Tutor

The Carnegie Mellon University (CMU) Lisp Tutor is a highly developed ITS that helps teach Lisp to college students. Two Lisp Tutor evaluations have estimated its instructional effectiveness. These evaluations are important, and will be reviewed in greatest detail, because they set the pace for many of the trends and findings that appear in other ITS evaluations.

The Lisp Tutor is used in a CMU Lisp course and follows the course text. The ITS is divided into 16 lessons that correspond to the 16 chapters contained in the course text, Essential Lisp

(Anderson, Corbett & Reiser, 1987). The Lisp Tutor was initially designed to cover the material presented in the first chapters of the Lisp course text and was slowly expanded to encompass the remaining course material.

The first evaluation study utilized three groups of ten subjects: a Lisp Tutor condition, a human tutor condition, and a self-taught condition (Anderson & Reiser, 1985; Anderson, Boyle, & Reiser, 1985). During the first evaluation, the Lisp Tutor covered 6 chapters in the course text, which represents 38 percent of the course. Dependent variables included an estimate of the time spent learning Lisp and the score of the individuals on a Lisp classroom exam, i.e., a power test. Analyses of the performance data indicated that the groups differed in time to master the lessons; the human tutor, ITS, and self-taught conditions required 11.4, 15, and 26.5 hours to cover the same material. The classroom test scores verified that the three groups were equivalent in their mastery of the material.

The second study (Anderson & Reiser, 1985) utilized two groups of 10 subjects: an ITS condition, and a traditional group instruction condition. At the time the second study was undertaken, the Lisp Tutor covered the material in 9 of the 16 chapters contained in the CMU Lisp course, which represents 56 percent of the course. The data analyses indicate a 30 percent savings in time for the ITS group along with a 43 percent better score on a knowledge test following the 9th chapter.

One noteworthy aspect of the first Lisp Tutor evaluation is the three group design: an ITS group, a human tutor group, and a traditional classroom group. Most ITS evaluations have not included a human tutor condition. Its inclusion allows the quality of the ITS instruction to be assessed relative to the human tutor interactions as well as relative to traditional classroom instruction.

An important feature of the Lisp ITS evaluation is the scope of the system: the ITS is designed to affect students' performance over a large proportion of the course. This feature is important given the small sample sizes used in the study. An ITS should have a wide scope in order to consistently demonstrate significant improvements in performance.

A notable trend in the Lisp evaluations is that the effect of the Lisp tutor on performance is not always evident. The Lisp Tutor evaluation indicated group differences on material covered in later chapters; but differences were not demonstrated for the material covered in the early chapters. One explanation for the lack of differences in early course material is that past individual differences in general programming knowledge result in increased variance on course exams and make it difficult to demonstrate group differences on general programming topics.

The effect of increased variance is apparent because of the small sample sizes used in the two studies. This is demonstrated by analyzing the power of the study. Assuming that the ITS produces an effect size of 1.0 standard deviation, [ES=(mean_1-mean_2)/sd_common] in test performance and that the alpha statistic is set to .05 (2-tail), a sample size of 10 implies that the likelihood of demonstrating a group difference is only 56 percent (Cohen, 1977). The power analysis underscores the importance of designing an ITS to have a major impact on student performance. Anderson was lucky to have demonstrated group differences.

Although the effect of previous general programming knowledge would be to increase group variance throughout the course, this effect would be larger for material covering general topics that correspond to other programming languages and courses. It follows that group differences are more likely to be demonstrated for course topics that are unique to Lisp. In fact, the Lisp Tutor evaluations report that group differences only appeared for topics covered later in the course, for example recursion, and not for the course material covered early in the course and that span other programming languages, e.g., iteration (Anderson & Reiser, 1985). This pattern is consistent with the variance explanation; topics such as recursion are less relevant to other programming languages and are less likely to be influenced by previous individual differences in programming knowledge and skills than are general topics such as iteration. This effect could have been partially controlled by including a covariate quantifying general programming knowledge; however, covariates were not used.

One problem with both Lisp Tutor evaluation reports is that they contain very few descriptive statistics. Estimates of variance, test reliability, and effect sizes are not reported. Sample size is reported in only one of the two articles describing the Lisp Tutor evaluations.

Proust

Proust is designed to assist students learning Pascal by tutoring them to solve two programming problems. The diagnostic power of Proust was first validated by determining the proportion of answers to the two programming problems that could be correctly diagnosed. When two hundred and six problem answers were submitted to Proust, Proust correctly diagnosed 79 percent of the programs and identified 94 percent of the bugs (Johnson &

Soloway, 1985). Thus, although Proust's scope is limited to two problems, it does understand most of the mistakes that students make while solving the problems.

The system was evaluated at Yale University in a value-added design (Center for the Study of Evaluation, 1986). In the Pascal course at Yale, students are given class lectures and are assigned eight optional homework problems. Two of the homework assignments correspond to the two problems supported by Proust.

Two groups were experimentally produced for the Yale evaluation. The first group was given access to Proust while the second group was denied access. Because the Yale homework was optional, students were not required to complete either homework assignment. Thus the evaluation followed a two by four group design in which the first dimension corresponded to access to Proust while the second dimension corresponded to the completion of the first, second, both or neither home work assignment. The design, and sample size for each cell, are contained in Table 1.

Table 1

Proust Evaluation Design

Opportunity	Home Work Completed				Total
	First	Second	Both	Neither	
Proust	12	2	19	36	69
No Proust	15	4	19	34	72
Total	27	6	38	70	141

Proust outcome measures consisted of student performance on the midterm and final exams, and student survey ratings. Covariates, which were used in the Proust evaluation, consist of measures such as grade point average and Scholastic Aptitude Test score, as well as survey data collected at the beginning of the term.

The data analyses indicate main effects across the homework condition, but no effect for tutorial opportunity. The objective data did not support the expectation that Proust would improve performance in the Yale Pascal programming classes.

There are several problems with the Proust design. First, the ITS could have only a minimal impact on course performance because the ITS was designed to cover a trivial portion of the material in the Yale course, two optional problems. This design is equivalent to estimating the impact of human tutors on student final exam performance by allowing students to be tutored on two of eight optional homework assignments. Even if differences had been demonstrated, they would have been small compared to the effect size of 2.0 standard deviations claimed for human tutoring.

A second experimental design problem is that by utilizing optional homework assignments, tremendous self selection is allowed in the decision of whether or not to complete the homework. Even if significant results had been obtained, explaining the results would have been difficult. Any actual differences could have been argued to reflect self-selection.

A third problem with the Proust evaluation is the small sample size. Only 33 students used the ITS for either homework assignment and only 19 students used it for both. The effect of using a small sample size is that only moderately large effects have a reasonable probability of being demonstrated. Cohen (1977) was used to estimate the power of a research design given an effect size estimate and the available sample size. The tables in Cohen were consulted under the assumption that alpha would be set to .05 (2-tail) and that the effect size of the experimental condition was .50 standard deviations. A sample size of 33 per group, which equals the number of students who used Proust for either homework assignment, corresponds to a power estimate of 51 percent; that is, assuming an effect size of .50 standard deviations, the probability of obtaining significant group differences is 51 percent. A probability of 32 percent is obtained for a sample size of 19, which corresponds to the number of students who used the system for both assignments.

However, the two power estimates overestimate the potential power of the Proust evaluation design because a .50 effect size estimate is optimistic. It is unlikely that human tutors, who are available for one or two of eight optional homework assignments, could improve midterm or final performance by .50 standard deviations; a smaller effect would be expected for an ITS.

An additional problem with the Proust design is that it utilized only two conditions: a group instruction condition and an ITS condition. Even if significant differences had been obtained, it would not have been possible to compare the effectiveness of Proust to the effectiveness of human tutors. Without this comparison it would not be possible to determine the extent to which Proust approximates human tutorial effectiveness.

Had the Proust evaluation used a subset of questions that corresponded to the two home work problems supported by the ITS, then the evaluation might have indicated improvement for that material. By focusing on global performance measures, such as midterm and final exam scores, the existence of an effect due to Proust became difficult to demonstrate. These problems indicate the importance of utilizing outcome variables that measure performance at various levels of transfer.

West

The West ITS helps children play the game, "How the West was Won". How the West was Won is designed to improve basic mathematical skills by requiring children to solve a large number of mathematical computations while they play.

The West evaluation utilized three groups: a West control group, a math plus West group, and a math plus strategy plus West group (Center for the Study of Evaluation, 1986). The sample sizes corresponding to each of the three groups were low: 5, 6, and 7. This study evaluated the effect of the math-key and strategy-key as an adjunct to West.

Students were obtained from fourth, fifth, and sixth grade classrooms, interviewed about attitudes towards games, and given a Math Pretest. The game was then described to the students and they were given a math workbook or a math plus strategy workbook depending upon their experimental condition. The students then worked in the appropriate workbook for 20 minutes. The next day subjects were allowed to play West for 45 minutes. The following day, the students were tested on math and strategy posttests and were interviewed individually about West.

Group differences were not demonstrated for any of the dependent variables. The only statistically significant finding in this study was that student performance on the math test improved from the pretest to the posttest. However, the magnitude of the effect can not be estimated because variance estimates were not reported. The importance of this effect is difficult to assess because there were no group differences.

This study is confounded by many of the same problems noted in the Lisp Tutor and Proust evaluations. The study used small samples composed of subjects from different grade levels, a method that is sure to increase group variance. Performance was assessed with a general math test after the subjects were in their experimental learning condition for less than one hour on material with which some of them may have already been familiar; it is doubtful that any tutoring system would have had a large effect after one hour of tutoring. Finally, descriptive statistics are not reported.

In contrast to the Proust evaluation, which used global effectiveness measures, the West evaluation used narrow measures. Ideally, these two strategies should be combined to accurately estimate the narrow and global impact of the ITS.

MACH III

The Maintenance Aid Computer for the HAWK - Intelligent Institutional Instructor (MACH III) was designed to train novice mechanics to troubleshoot the HAWK radar system. There are two main components to the ITS: the transmitter portion and the receiver portion. When the system was evaluated, the MACH-III could be used extensively during the transmitter portion of the course, but for less than one hour during the receiver portion of the course.

The evaluation (Kurland, Granville, & MacLaughlin, 1990) compared 2 groups of 11 students. The control group received traditional group instruction while the ITS group was given access to MACH-III for the two sections of the course supported by the system. Access to the ITS was integrated into the course without changing the total course length. The ITS was used for approximately four days of instruction.

Effectiveness measures were obtained on the basis of written and practical exams for the course. The evaluation focused on those exams that were directly relevant to either the transmitter or the receiver portion of the course. Students and instructors were extensively interviewed to identify possible improvements in the ITS and to determine how the ITS could be most effectively used.

A one standard deviation group difference was demonstrated for the transmitter written exam. However, group differences were not demonstrated on either the receiver written exam, the receiver practical exam, or the transmitter practical exam.

The group difference on the transmitter written exam is consistent with the large scope of the ITS for transmitter topics. The ITS was used for approximately one-third of the two and one-half weeks of the course devoted to the receiver and transmitter portions of the course. On this exam, the difference between the two groups was significant and was equal to approximately one standard deviation.

The lack of significant differences on the two practical exams reflects a ceiling effect on these exams. The mean scores of the two groups across the two sets of practical exams range from 96.7 to 99.6 percent, (standard deviations are not reported). Ceiling effects did not occur for the written exams; these means ranged from 78.6 to 89.1 percent.

The lack of a group difference on the receiver written exam is consistent with the small extent to which the ITS was used for receiver topics, i.e., less than one hour. As noted in the critiques of other ITS evaluations, it is difficult to demonstrate the effectiveness of an ITS that has a narrow scope. One would certainly not attempt to demonstrate improvement on a course exam after substituting a one hour block of human tutorial instruction for traditional classroom instruction. Why expect an effect for an ITS?

The major weakness with the MACH-III evaluation was the non-inclusion of a human tutorial condition. If a human tutorial group had been included in the evaluation, then it may have performed equivalently to the ITS group, i.e., one standard deviation above the mean performance of the traditional group. Such a finding would imply that the ITS and human tutors were approximately equivalent in effectiveness. If the human tutor group had performed better than both groups, e.g., two standard deviations above the control group mean, then the data would have suggested that the ITS could be made still more effective by emulating student-tutor interactions.

Smithtown

Smithtown (Raghavan & Katz, 1989) is designed to assist students in introductory micro-economics courses. The ITS covers approximately one-third of the topics included in an introductory economics course at the University of Pittsburgh.

Smithtown was evaluated in a substitution design, with thirty subjects. This evaluation compared students who learned economics with Smithtown to students who learned economics in a standard introductory economics course. The data indicate that students require 5 hours of interaction with Smithtown to learn an amount equivalent to students who spend 12 hours (plus homework) in a standard lecture based class. However, the evaluation report fails to report test scores that would prove equivalent levels of learning.

A second evaluation included a third group of subjects, who attended the class lectures and utilized Smithtown. The report of this evaluation states that the third group of subjects learned at a rate that surpassed the other two groups. No additional information or descriptive statistics are reported for either of these evaluations; covariates are not discussed.

This research is difficult to evaluate because few statistics are reported. The results are interpreted by Raghavan and Katz to indicate that the ITS utilized student time more effectively than did classroom instruction. However, this conclusion is warranted only under the assumption that the groups performed equivalently on the course tests. Because the course tests are not adequately described, it cannot be speculated whether this assumption holds. Nonetheless, the ITS covers a large portion of the course, addresses an educational need, and appears to be supported by the evaluation.

Prewriting Tutor

The Prewriting Tutor was designed to assist students during the prewriting stage of composition (Gillis, 1984). The tutor supports approximately five hours of instruction. The Prewriting Tutor was evaluated by comparing the effectiveness of the ITS with a human tutor condition and a traditional group instruction condition across several dependent measures that quantified different aspects of the prewriting process. Although each group contained up to 50 individuals, some of the dependent measures were difficult to obtain and full data sets were available only for between 12 and 20 individuals for each group. Differences favoring the ITS condition were demonstrated for most of the dependent measures.

The Prewriting Tutor evaluation has many of the same features as the Lisp ITS evaluations; three experimental conditions were used and positive effects were demonstrated for the tutor. One difference between the Prewriting tutor and other effective tutors, i.e., Lisp ITS, MACH III, and Smithtown, is that the Prewriting Tutor supported much less instruction then these other ITSs - only five hours. Consistent with the smaller scope of the Prewriting tutor, the effect size of the tutor was approximately one-half standard deviation, much smaller than the effects demonstrated for the other effective tutors.

Pixie

The Pixie tutor is being developed to support algebra instruction. Two reports by Sleeman, Kelly, Martinak, Ward and Moore (1988, 1989) summarize a series of four experiments performed in relation to the Pixie project.

The experiments were designed to assess the relative effectiveness of three different approaches to tutoring: Model Based Remediation, Reteaching, and the control condition. In the first condition, Model Based Remediation (MBR), students were given tutorial instruction based on past student performance. This approach requires student models, and was expected to improve student performance by focusing on critical aspects of learning. This approach is oriented around the capabilities of an ITS. In the reteaching condition, students repeated a segment of instruction after an error was committed. No attempt was made to orient the instruction to the student in this condition. Reteaching is well suited as a CAI oriented tutorial approach because it does not require the system to contain a student or expert model to analyze answers to problems. In the control condition, errors were pointed out to the students, but remediation was not given.

Evaluation of the three tutorial approaches was undertaken on students who had below average achievement in algebra classes. In the experimental paradigm, students were placed in one of the three experimental conditions and worked on algebra problems for approximately a one-week period. A test was administered to the students at the end of the intervention period to estimate their gain in algebra knowledge and to determine group differences in achievement.

The first experiment compared the performance of students who were given access to the three tutorial approaches via microcomputers. A class of 24 below average students were divided into three groups of 8 students. No differences were demonstrated between the two experimental conditions, although both conditions performed significantly better on the test than the control condition in which students were simply informed when errors were made (Sleeman et al., 1988).

A second experiment was identical to the first, but human tutors were used instead of microcomputers because the first experiment was interpreted as indicating that the computers were not effectively tutoring. Although both groups performed better than the control group on the outcome measure, again there were no differences between the two experimental conditions (Sleeman et al., 1988; Sleeman et al., 1989).

A third experiment was designed to determine why the two experimental conditions did not differ in the first two experiments. The third experiment used human tutors to estimate the difference in effectiveness between the two original experimental conditions and two enhancements of the MBR tutorial approach. Again, no differences were demonstrated between any of the four experimental conditions (Sleeman et al., 1988; Sleeman et al., 1989).

A fourth experiment used two pretests to identify errors that subjects were consistently making and provide human tutoring for these errors. The fourth experiment utilized 25 students that were divided into an MBR group (n=9), a Reteaching group (n=8), and a control group (n=8). Again, no differences were found between the two experimental groups. The only significant differences were an improvement from pretest to posttest for both groups and a difference between both experimental conditions and the control on the posttest (Sleeman et al., 1988).

The problems associated with the Pixie research are similar to those found in other ITS evaluations. First, all the Pixie experiments used small sample sizes of subjects. However, only large effect sizes can be detected with small sample sizes. It is not surprising that the experimental conditions did not differ because there is no evidence to suggest that one week of human tutoring can have a large impact on performance. Furthermore it is difficult to interpret the reported gain scores because variance estimates are not provided.

Second, the Pixie report provides very few numeric statistics. Estimates of variance, test reliability, effect size, and inferential statistics are not reported.

Third, the meaning of any comparison with the control group is questionable because the control condition does not correspond to a tutorial system that would be used. A more meaningful comparison would involve a traditional instruction condition, but this group was not included.

One important finding of these studies is the lack of differences between the MBR and Reteaching conditions. The authors indicate that these groups are analogous to an human tutor and a conventional approach. The fact that no differences were demonstrated implies that human tutors may be no more effective than traditional group instruction. The Pixie findings underscore the importance of including a human tutoring group, and a traditional classroom instructional group, in an ITS effectiveness evaluation.

One advantage to the Pixie evaluation design is that it demonstrates that aspects of the ITS may be manipulated to determine their internal effectiveness and assess their relative importance. The Adaptive Computerized Training System, which is described next, also utilized this approach. This strategy could be used to determine the instructional impact of ITS components such as high bandwidth student diagnostic routines and on line coaching utilities.

Adaptive Computerized Training System (ACTS)

ACTS was one of the first ITSs and was designed to support electronics troubleshooting. Crooks, Kuppin, and Freedy (1978) describe an evaluation of ACTS that compared three groups of students who were given access to the ITS. The study was designed to assess the impact of providing various amounts of coaching on student performance. Although this study was not designed specifically to evaluate the effectiveness of the ITS, the evaluation did address the effectiveness of some of the components of the ITS and has the same limitations found in the other ITS evaluations.

The ACTS presents students with problems and records the number of decisions that the student makes in order to solve the problem. The system also estimates the cost of the solution followed by the student. These two values were used as outcome measures in the evaluation of the system.

ACTS provides students with two types of on-line help. The help function can be used be used to obtain a list of recommended alternatives for the student to test. The system will also provide feedback to the student after a test point has been selected. Parts of the help-function were disabled across the three experimental conditions for the ACTS evaluation.

In the full access condition, students could use the help utility to obtain a list of three recommended alternatives to test. After choosing an alternative, the students were given feedback describing the choice that the computer would have made. In the second condition, the students were given feedback by the computer after making a decision; but the students were not allowed to obtain a list of recommended actions. In the third condition, students were not given access to the recommended actions or feedback about their own decision.

The design utilized three groups of two students. Subjects were first given an introductory session of five troubleshooting problems with neither the on-line help nor feedback from the computer. The six students were then placed in the three experimental conditions and used the system for three sessions to assess the impact on performance.

Access to feedback and the help utility were varied over the three sessions. The first group was given access to the help utility and feedback during the fist session, only feedback for the second session, and denied access to both utilities during the third session. The second group was given feedback during the first two sessions, and denied access to both utilities during the third session. The control group was not given access to the help utility or feedback during any of the three sessions. The third session can be viewed as a test session because the subjects were all treated equivalently.

Only small differences were demonstrated between the three groups on either dependent measure. No significant group differences were reported.

One problem with this evaluation is that the sample sizes were inadequate for even a large effect size. Furthermore, had inferential statistics indicated group differences, the differences would be difficult to interpret because they could reflect differences in aptitude or knowledge across the six students.

A second problem is that a single session with an ITS is unlikely to produce a measurable long-term effect on performance. Even if a sufficient sample size had been available, it is unlikely that one or two sessions would have been sufficient to demonstrate an effect on performance.

A third problem is that only circumscribed interpretations can be based on comparisons with the control group, which was denied access to any tutorial system and was essentially a self-taught group. On the assumption that the ITS was evaluated with a well designed study, then a lack of significant group differences between an ITS group and a self-taught group would indicate that the ITS was a failure. For an ITS to be a success, it should be at least as effective as traditional classroom instruction, which can be assumed to be more effective than self-study. To find that an ITS group performs better than a self-taught group would be a very limited validation of a training technology that attempts to surpass traditional training methods by emulating one-on-one human tutor interactions.

ITS Evaluation Criteria

This review points to three criteria that are particularly relevant to ITS evaluation research. First, ITS evaluations should utilize three experimental conditions: the ITS group, a traditional instruction group, and a human tutor group. Second, only well developed ITS applications that have an extensive scope should be used to estimate the impact of this technology on learning. Third, ITS evaluations should utilize larger group sizes than have been used in the past. This section argues for adopting these criteria for future ITS effectiveness evaluations. A discussion of several minor evaluation considerations is included at the end of this section.

Required Conditions: ITS, Human Tutor, & Traditional Instruction

Logic dictates that a minimum of three experimental conditions be included to evaluate the ITS effectiveness: a traditional classroom control, a human tutorial, and an ITS group. Although most ITS evaluations have not included these three groups, all three are critical in order to insure meaningful results.

ITS proponents have consistently justified ITS development on the claim that human tutoring is a highly effective means of instruction (Anderson & Reiser, 1985; Littman & Soloway, 1988). However, not all learning objectives should be assumed to be effectively taught through human tutoring. Comparing the effectiveness of a human tutor and a traditional instruction group would assess this assumption for the ITS subject matter and could justify basing ITS instruction on models of human tutoring. If human tutors cannot be demonstrated to be more effective than classroom instruction for that specific subject matter, then it is not logical to construct a system that emulates human tutors. Furthermore, human tutoring is not monolithic and can vary in components and approach; therefore, verifying the effectiveness of the modelled human tutoring approach is important. It may be preferable compare human tutorial approaches prior to developing the ITS.

If it can be assumed that human tutoring is more effective than traditional instruction, then including both controls in the evaluation allows the power of the research design to be assessed. Under this assumption, the experimental finding that the two groups do not differ is a Type II error and indicates problems with the experimental design. Such problems could occur if the evaluation used small sample sizes or inaccurate performance measures.

If the human tutor and the traditional instruction groups differ, then the magnitude of that difference can be used to estimate the expected magnitude of the differences between the ITS and traditional instruction groups. This estimate may be used to assess the power of the research design used for the evaluation; this is particularly important if the evaluation does not demonstrate differences favoring the ITS group.

A comparison between human tutors and ITS is also important to estimate the extent to which pedagogically important aspects of human tutor interactions have been encoded into the ITS. It can be expected that as ITS technologies improve, computer based tutors will approximate the effectiveness of human tutors. Comparison with human tutors allows ITS developers to estimate the quality of their programs and determine when other models are needed to further improve the ITS.

Evaluating the cost-effectiveness of an ITS requires the ITS group to be compared to a traditional classroom control because it is usually the least expensive instructional alternative.

(Although, if classroom instructors are not required conventional CAI may b: the least expensive instructional medium.) If ITS technology cannot be shown to be either more effective than traditional instruction, or equally effective and less costly than traditional instruction, then it is not logical to develop these systems for practical applications.

Impact of the System

ITS evaluations have assessed the impact on learning of both extensive and narrow ITS applications. Extensive systems are defined as systems that can be expected to have a large impact on performance by virtue of their wide scope, while a small impact on performance can be anticipated for narrow systems.

The magnitude of impact is important for two reasons. First, demonstrating a small impact requires much greater evaluation resources than are needed to demonstrate a large impact. Second, whether or not adequate evaluation resources are available, an ITS must be extensive in order to produce a large impact on performance. An ITS that has only a small impact on performance cannot justify continued ITS research and development. It follows that ITS evaluations should focus on extensive tutors.

Evaluations of Narrow ITS Applications. Evaluations performed on narrow ITS applications have not been supportive of this technology. This problem is most apparent in the evaluations of older ITSs and probably reflects the fact that these applications were designed to demonstrate ITS instructional capabilities. Less attention was paid to either the scope or the practical impact of the system. The Proust, West, and original Pixie work exemplify these applications.

These systems were evaluated after they had been used for between one and five hours. The problem with evaluating narrow systems is most apparent in the Pixie research, which failed to demonstrate differences regardless of whether computers or human tutors were used to teach students. The Pixie research suggests that group differences would not have been demonstrated had human tutors been used instead of the computers for the Proust and West evaluations. This is because the time allotted to tutoring was less for the West and Proust ITS evaluations than for the Pixie evaluation and can therefore be expected to have a small impact. Given the small sample sizes used in these evaluations, the lack of empirical differences should not influence expectations concerning the effectiveness of ITS technology.

Evaluations of Extensive ITS Applications. In contrast to the equivocal evaluation data for narrow ITS applications, data collected for extensive applications, developed for practical use, support continued ITS research and development. The MACH III, Lisp Tutor, and Smithtown exemplify extensive systems. All three tutors were designed for an actual training requirement and cover a large amount of course material. The MACH III supports 32 hours of Army training, the Lisp Tutor covers a one semester course at Carnegie Mellon University, and Smithtown corresponds to approximately one-third of an economics course at the University of Pittsburgh.

The evaluations of these three ITS applications demonstrated statistically significant group differences. Furthermore, the group differences are meaningful in that each of the tutors had a substantial impact on course performance. The effects range from a 58 percent savings in learning time for Smithtown students to a 1.0 standard deviation improvement on test scores for soldiers taught with the MACH III. The Lisp Tutor evaluation data are consistent with these estimates.

Need for the Evaluation of Extensive Systems. The ITS evaluations suggest that only extensive ITS applications are favorably evaluated. The tutorial systems can be placed on a continuum ranging from narrow to extensive. The successful evaluations were performed on applications that covered at least one-third of the material in a typical college class or over 30 hours of material in an Army course. The non-successful evaluations were conducted on systems that supported less than five hours of instruction. These values can be used to categorize the scope of an ITS as either extensive, narrow, or borderline. The boundaries between these categories should be viewed as flexible and are doubtlessly dependent upon a number of factors apart from the scope of the tutor, e.g., student variance in prior knowledge and intelligence.

A practical advantage to designing ITS applications for realistic instructional objectives is that alternatives to ITS instruction already exist and can be used to construct control conditions in the evaluation. On the other hand, traditional classroom instruction and human tutor modules will not be readily available for the types of topics addressed by narrow systems. Instead, the evaluator must design the traditional instruction, as well as the other two conditions. This entails an obvious conflict of interest, which may have biased CAI evaluation studies (Clark, 1985) and could be problematic to ITS evaluations. This problem is less likely to occur when an ITS application is developed for a realistic application where traditional instruction is available.

Sample Size Requirement: Power Analyses

One problem with many ITS evaluations is the use of small groups of subjects. Table 2 contains a listing of sample sizes for evaluations of eight ITS applications and summarizes their outcome. Table 2 shows that all studies used sample sizes that were less than 20. This is problematic from the standpoint of power analyses because small group sizes result in very weak ITS evaluation designs. Much larger numbers of subjects need to be included in ITS evaluation designs in order to accurately assess the impact of ITS technology.

The data summarized in Table 2 demonstrate that group size and effect size are linked to the probability that the system is likely to have a demonstrable effect on performance. The only favorable evaluation studies were those that utilized larger samples and those that have already been identified as having a substantial impact on performance.

Table 2

Sample Sizes Used in ITS Evaluations

ITS	Sample Size Per Group	Number of Conditions	Outcome[a]	Reference[b]
Lisp ITS	10	3	1	Anderson 1985
Lisp ITS	10	2	1	Anderson 1985
Proust	2 - 19	2[c]	0	Center 1986
West	5 - 7	3	0	Center 1986
ADCS	2	3	0	Crooks 1978
Prewriting	12 - 20	3	1	Gillis 1983
MACH III	11	2	1	Kurland 1990
Smithtown	15	2	1	Raghavan 1989
Smithtown	15	3	1	Raghavan 1989
Pixie	8	3	0	Sleeman 1989
Pixie	19	3	0	Sleeman 1988
Pixie	12	4	0	Sleeman 1988
Pixie	8 - 9	3	0	Sleeman 1989

a: "1" indicates group differences were demonstrated.
"0" indicates that group differences were not demonstrated.
b: References were abbreviated; refer to text.
c: Four levels within each condition.

The data in Table 2 reflect the mathematical relationship between sample size, power, alpha probability level, and effect size. The power of the study will increase with corresponding increases in either the effect size of the treatment or the sample sizes used in the research. Cohen (1977) has calculated power estimates for various effect sizes and sample sizes. A portion of the values reported by Cohen is listed in Table 3.

Assuming an effect size of 1.0 standard deviation unit, which is the value estimated for the MACH III effect size and is consistent with the Lisp Tutor and Smithtown effect size, Table 3 indicates that the sample sizes used in these three successful evaluations were marginally adequate. The power associated with these studies is approximately .55, .60, and .83 for the sample sizes used in the Lisp Tutor, MACH III, and Smithtown evaluations. In fact, the designers of these evaluations were fortunate to have demonstrated group differences in these studies.

Table 3

Power Analysis Estimates: Required Sample Size Per Group at the alpha=.05 (2-tail) Level by Effect Size and Power

Power[1]	Effect Size=¦MeanA-MeanB¦/sigma						
	.20	.40	.60	.80	1.00	1.20	1.40
.25	84	22	10	6	5	4	3
.50	193	49	22	13	9	7	5
.60	246	62	28	16	11	8	6
.80	393	99	45	26	17	12	9
.95	651	163	73	42	27	19	14
.99	920	231	103	58	38	27	20

[1] From Statistical Power Analysis for the Behavioral Sciences Revised Edition (p. 55) by J. Cohen, 1977, New York: Academic Press. Copyright 1977 by Academic Press. Adapted by permission.

The impact of effect size on power is also demonstrated by comparisons within the MACH III and Lisp Tutor evaluations. The MACH III evaluation compared experimental and control groups on four separate exams. Three of the exams were related to a small portion of the ITS or were psychometrically poor; group differences were not observed for these scales. On the fourth exam, the groups were significantly different at the p=.04. Thus the MACH III evaluation nearly led to a finding of no significant group differences.

The Lisp Tutor was evaluated twice, the ITS covered 38 percent of the course for the first evaluation and 56 percent of the course during the second evaluation. Although the ITS groups in both evaluations spent less time solving problems, the ITS group performed significantly better than the control group only on the knowledge exam used in the second evaluation. In the first evaluation, group differences were not demonstrated on the knowledge exam. It follows that, given the small sample sizes used in the Lisp evaluations, the effect of the ITS on student performance could have been easily missed.

These considerations underscore the importance of evaluating extensive ITS applications with large sample sizes. Using a conservative power estimate of .80, sample sizes greater than 42 and 26 per group are required to avoid a Type 2 error, i.e., missing of a real effect, at the 95 and 80 percent probability levels. Refer to Table 3.

Other Considerations

Outcome measures. Ideally, performance data should measure time savings and power (percent correct). Unfortunately, these two categories conflict with each other because they are negatively correlated within individuals, i.e., the faster a person responds, the less likely that response is to be correct. Past ITS (and CAI) evaluations have used both types of measures, but the emphasis has usually been on power scales.

Attempting to demonstrate both time savings and test improvement can be problematic because the effect of the ITS is split across several dimensions. (Whereas by holding one dimension constant, e.g., time, the treatment effect would primarily affect the other dimension, i.e., percent correct.) Once the effect is split across several dimensions, demonstrating significant improvements due to the ITS on any one dimension may require a larger sample size. For this reason it is usually logical to limit variations in time on the ITS and concentrate on demonstrating differences on power scales.

Ideally, the tests should estimate the effectiveness of the ITS across several points of generalization. For example, the outcome measures should assess performance on problems specifically related to the tutor, as well as on general class exams. This strategy helps to insure that the evaluation will assess the overall effectiveness of the tutor and yield interpretable results.

Reporting Statistics. One shortcoming with many ITS evaluations is that the tests are not adequately described. Neither reliability, validity, effect size, nor variance estimates are reported. Covariates are often neither included nor discussed in the evaluation reports. Without these values it is difficult to assess the evaluation. Including these values is particularly important when the group differences are not significantly different. In this case it is critical that the power of the research design be estimated.

Summary

This paper reviewed standards and procedures that have been used to evaluate intelligent tutoring systems. On the basis of the review, three criteria are proposed to evaluate the external effectiveness of intelligent tutoring systems. First, performance data should compare the instructional effectiveness of the intelligent tutoring system, human tutors, and traditional group instruction. Second, only extensive ITS applications should be evaluated. Third, large groups of students are required for these evaluations. By adapting these criteria, results of ITS evaluations can be used to guide future ITS application development.

References

Anderson, J. R., Boyle, C. F. & Reiser, B. J. (1985). Intelligent Tutoring Systems, Science, 228, 456-462.

Anderson, J. R., Corbett, A. T., & Reiser, B. J. (1987). Essential Lisp. Reading, MA: Addison-Wesley Publishing Company.

Anderson, J. R. & Reiser, B. J. (1985). The Lisp tutor. Byte, 10 (4), 159-175.

Bangert-Downs, R. L., Kulik, J. A. & Kulik, C. C. (1985). Effectiveness of computer based education in secondary schools. Journal of Computer Based Instruction, 12, 59-68.

Bloom, B. S., (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring, Educational Researcher, 13 (3), 4-16.

Center for the Study of Evaluation, University of California, Los Angeles. (1986). Intelligent computer aided instruction (ICAI): Formative evaluation of two systems (Army Research Institute Research Note 86-29). Alexandria, VA: U.S. Army Research Institute. (DTIC No. AD-A167 910)

Clark, R. E. (1985). Evidence for confounding in computer-based instruction studies: Analyzing the meta-analyses. Educational Communication Technology Journal, 33, 249-262.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York, NY: Academic Press.

Crooks, W. H., Kuppin M. A. & Freedy, A. (1978). Application of adaptive aiding systems to computer-assisted instruction: Adaptive computer training system (ACTS) (Army Research Institute Technical Report 78-A6). Alexandria, VA: U.S. Army Research Institute. (DTIC No. AD-A056 900).

Gillis, P. D. (1983). The Utilization of Computer Technology as a Means of Teaching and Evaluating Prewriting Processes. Unpublished doctoral dissertation, Georgia State University, Atlanta, GA.

Johnson, W. L. & Soloway, E. (1985). Proust. Byte, 10 (4), 179-190.

Kulik, J. A., Bangert, R. L. & Williams, G. W. (1983). Effects of computer-based teaching on secondary school students. Journal of Educational Psychology, 75, 18-26.

Kulik, C. C., Kulik, J. A. & Shwalb, B. J. (1986). The effectiveness of computer based adult education: A meta-analysis. Journal of Educational Computing Research, 2, 235-252.

Kurland, L. C., Granville, R. A., & MacLaughlin, D. M. (1990). Design, development, and implementation of an intelligent tutoring system (ITS) for training radar mechanics to troubleshoot. Unpublished manuscript, Bolt, Beranek, and Newman, Systems and Technologies Corporation, Cambridge, MA. Littman, D., & Soloway, E. (1988). Evaluating ITSs: The cognitive science perspective. In M. C. Polson & J. J. Richardson (Eds.), Foundations in intelligent tutoring systems (pp. 209-242). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Priest, T., & Young, R. (1988). Methods of evaluating micro-theory systems. In J. Self (Ed.), Artificial Intelligence and Human Learning: Intelligent Computer-aided Instruction, London: Chapman and Hall Computing.

Orey, M., Legree, P., Gillis, P, & Bloom, E. (1991). Evaluation of Intelligent Tutoring Systems: What's been done and what can be done. Proceedings of the 9'th Annual Conference of the Association for Educational Communications Technology.

Raghavan, K. & Katz, A. (1989). Smithtown: An Intelligent Tutoring System. Technological Horizons in Education Journal, 17 (1), 50-54.

Sleeman, D., Kelly, A. E., Martinak, R. Ward, R. D., & Moore, J. L. (1988). Diagnosis and remediation in the context of intelligent tutoring systems (ARI Research Note 88-66). Alexandria, VA: U.S. Army Research Institute. (DTIC No. AD-A199 024).

Sleeman, D., Kelly, A. E., Martinak, R. Ward, R. D., & Moore, J. L. (1989). Studies in the diagnosis and remediation of high school algebra students. Cognitive Science, 13, 551-568.